Adaptive warped kernel estimators

Gaëlle Chagny

ADAPTIVE WARPED KERNEL ESTIMATORS

GAËLLE CHAGNY[A] *

Abstract. In this work, we develop a method of adaptive nonparametric estimation, based on "warped" kernels. The aim is to estimate a real-valued function s from a sample of random couples (X, Y). We deal with transformed data $(\Phi(X), Y)$, with Φ a one-to-one function, to build a collection of kernel estimators. The data-driven bandwidth selection is done with a method inspired by Goldenshluger and Lepski (2011). The method permits to handle various problems such as additive and multiplicative regression, conditional density estimation, hazard rate estimation based on randomly right censored data, and cumulative distribution function estimation from current-status data. The interest is threefold. First, the squared-bias/variance trade-off is automatically realized. Next, non-asymptotic risk bounds are derived. Last, the estimator is easily computed thanks to its simple expression: a short simulation study is presented.

Keywords: Adaptive estimator. Censored data. Bandwidth selection. Nonparametric estimation. Regression. Warped kernel.

AMS Subject Classification 2010: 62G05; 62G08; 62N02.

1. Introduction

Let (X, Y) be a couple of real random variables, and $(X_i, Y_i)_{i=1,\dots,n}$ an *i.i.d.* sample drawn as (X, Y). The main goal of nonparametric estimation is to recover an unknown function s, linked with (X, Y), such as the regression function, from the data. Among the huge variety of methods that have been investigated, the use of transformed data $(F_X(X_i), Y_i)$, with F_X the cumulative distribution function (c.d.f.) of X, has received attention in the past decades. In this context or with similar tools, both kernel and projection estimators have been studied in random design regression estimation (Yang 1981, Stute 1984, Kerkyacharian and Picard 2004, Akritas 2005, Pham Ngoc 2009, Kulik and Raimondo 2009, Mammen *et al.* 2012, Chagny 2013a), conditional density or c.d.f estimation (Stute 1986, Mehra *et al.* 2000, Chagny 2013c), for the white noise model (Chesneau 2007) or to deal with dependent data (Chesneau and Willer 2012). However, to our knowledge, few papers focus on the problem of adaptivity of such "warped estimators". The aim of the present work is twofold: first, we want to show that a warping kernel device can be applied to various estimation problems, including survival analysis models (see examples below). Secondly, we address the problem of bandwidth selection, with the intention of providing an adaptive "warped" estimator, which satisfies nonasymptotic risk bounds.

The basic idea, which motivates the study of warped kernel estimators introduced by Yang (1981), can be first explained in the classical regression framework. Here, the target function is the conditional expectation, $s : x \mapsto \mathbb{E}[Y|X = x]$ *i.e.*

$$s(x) = \frac{1}{f_X(x)} \int_{\mathbb{R}} y f_{(X,Y)}(x, y) dy, \tag{1}$$

* Corresponding author. Email: gaelle.chagny@gmail.com

[A] *Laboratoire MAP5 (UMR CNRS 8145), Université Paris Descartes, and LMRS (UMR CNRS 6085), Université de Rouen, France.*

when a density $f_{(X,Y)}$ for the couple (X,Y) exists, and where f_X is the marginal density of the design X. Historical kernel methods were initiated by Nadaraya (1964) and Watson (1964). The famous estimator named after them is built as the ratio of a kernel estimator of the product sf_X divided by a kernel estimator of the density f_X:

$$\tilde{s}^{NW} : x \mapsto \frac{\frac{1}{n}\sum_{i=1}^{n} Y_i K_h(x - X_i)}{\frac{1}{n}\sum_{i=1}^{n} K_h(x - X_i)},$$

where $K_h : x \mapsto K(x/h)/h$, for $h > 0$, and $K : \mathbb{R} \to \mathbb{R}$ such that $\int_{\mathbb{R}} K(x)dx = 1$. Adaptive estimation then requires the automatic selection of the bandwidth h, and the ratio form of the estimate suggests that two such parameters should be selected: one for the numerator, and one for the denominator. From the theoretical point of view, there is no reason to choose the same. Nevertheless, nonasymptotic results such as oracle-inequality are difficult to derive for an estimator defined with two different data-driven smoothing parameters. See Penskaya (1995) for a thorough study of the ratio-form estimators. Moreover, when the design X is very irregular (for example when a "hole" occurs in the data), a ratio may lead to instability (see Pham Ngoc 2009). The warped kernel estimators introduced by Yang (1981) and Stute (1984) avoid the ratio-form. Indeed denote by $\hat{F}_n$ the empirical c.d.f. of the X_i's and let

$$\hat{s}_h = \frac{1}{n}\sum_{i=1}^{n} Y_i K_h\left(F_X(x) - F_X(X_i)\right), \text{ or } \hat{s}_h = \frac{1}{n}\sum_{i=1}^{n} Y_i K_h\left(\hat{F}_n(x) - \hat{F}_n(X_i)\right), \tag{2}$$

depending on whether the c.d.f. F_X is known or not. The following equality (see (8)) holds:

$$\mathbb{E}\left[Y K_h\left(u - F_X(X)\right)\right] = K_h \star (s \circ F_X^{-1})(u),$$

where $\star$ is the convolution product and $\circ$ is the composition symbol. Thus, the first estimator of (2) can be viewed as $\hat{s}_h = \widehat{s \circ F_X^{-1}} \circ F_X$. The main advantage is that its expression involves one bandwidth h only.

In this paper, we generalize the warping strategy to various functional estimation problems: as a first extension of (1), we propose to recover functions s of the form

$$s(x) = \frac{1}{\phi(x)} \int \theta(y) f_{(X,Y)}(x,y)dy, \tag{3}$$

for $\theta : \mathbb{R} \to \mathbb{R}$, and $\phi : \mathbb{R} \to \mathbb{R}_+\backslash\{0\}$. In this case, the warping device brings into play the transformation $(\Phi(X), Y)$ of the data, with $\Phi' = \phi$. The form (3) covers the additive regression model described above, by setting $\Phi = F_X$, and $\theta(y) = y$. But it also permits to deal with the simplified heteroskedastic model $Y = \sqrt{s(X)}\varepsilon$, where ε is an unobserved noise, centered, with variance equals to 1. In this case, $\Phi = F_X$, and $\theta(y) = y^2$.

In several examples however, the couple (X,Y) does not admit a density, but X admits a marginal density. Then (3) can be extended and the target function s takes the form:

$$s(x) = \frac{f_X(x)}{\phi(x)} \mathbb{E}[\theta(Y)|X = x]. \tag{4}$$

This allows to handle two classical settings in survival analysis: the interval censoring case 1, and right censored data. In the interval censoring model, case 1, the target function is $s(x) = \mathbb{P}(Z \leq x)$, where Z is a survival time, which is not observed, and we only know a current status at the observed time X of examination. We also know $Y = \mathbf{1}_{Z \leq X}$, which indicates whether Z occurs before X or not. We refer to Jewell and van der Laan (2004) for a review of the estimation methods in this setting (see also van de Geer 1993 for maximum likelihood estimation), and more recently to Ma and Kosorok (2006), Brunel and Comte (2009) or Plancade (2013) for investigations including adaptivity. In right-censored data, the function of interest at time x is the hazard rate function, that is the

risk of death at time x, given that the patient is alive until x. This model has been studied by Tanner and Wong (1983), Müller and Wang (1994) and Patil (1993), among all. Adaptive results are available for projection-type estimators (see Brunel and Comte 2005, 2008, Reynaud-Bouret 2006 or Akakpo and Durot 2010), but to our knowledge not for kernel estimators.

The paper is organized as follows. We present in Section 2 the estimation method, detail the examples illustrating the relevance of the introduction of a general target function s defined by (4). We also study the global risk of the warped kernel estimators with fixed bandwidths. Section 3 is devoted to adaptive estimation: we define a data-driven choice of the bandwidth, inspired by Goldenshluger and Lepski (2011) which allows to derive nonasymptotic results for the adaptive estimators. Oracle-type inequalities are provided for the M.I.S.E., and convergence rates are deduced under regularity assumptions. Sections 2 and 3 both deal with the case of known deformation Φ and with the case of an estimated deformation. In Section 4, the method is illustrated through numerical simulations. Proofs are gathered in Section 5. The supplementary material Chagny (2013b) is available for further details about technical computations.

2. Estimation method

2.1. **Warped kernel strategy.** Consider a sample $(X_i, Y_i)_{i=1,\dots,n}$ of *i.i.d.* random couples with values in $A \times B$, where A is an open interval of $\mathbb{R}$ and B a Borel subset of $\mathbb{R}$. We assume that X_i has a marginal density f_X and we aim at recovering a function $s : A \to \mathbb{R}$ linked with the distribution of (X_i, Y_i). To estimate s, we replace the explanatory variable X_i by $\Phi(X_i)$, where $\Phi : A \to \Phi(A) \subset \mathbb{R}$ is one-to-one and absolutely continuous. The data $(\Phi(X_i), Y_i)_{i=1,\dots,n}$ are called the warped sample with deformation function Φ. The sets $A, B, \Phi(A)$ are supposed to be given. The target function can be written as:

$$s(x) = g \circ \Phi(x) = g(\Phi(x)), \text{ with } g : \Phi(A) \to \mathbb{R}. \tag{5}$$

We first estimate the auxiliary function $g = s \circ \Phi^{-1}$ with Φ^{-1} the inverse function of Φ. In the general case, Φ is unknown and we must estimate it also. Let K be a function such that $\int_{\mathbb{R}} K(u)du = 1$ and set $K_h : u \mapsto K(u/h)/h$, for $h > 0$. We define, for $u \in \Phi(A)$,

$$\hat{g}_h(u) = \frac{1}{n} \sum_{i=1}^{n} \theta(Y_i) K_h \left(u - \hat{\Phi}_n(X_i) \right), \tag{6}$$

where $\theta : \mathbb{R} \to \mathbb{R}$ is a given function, $\hat{\Phi}_n$ is an empirical counterpart for Φ, and for $x \in A$

$$\hat{s}_h(x) = \hat{g}_h \circ \hat{\Phi}_n(x) = \frac{1}{n} \sum_{i=1}^{n} \theta(Y_i) K_h \left(\hat{\Phi}_n(x) - \hat{\Phi}_n(X_i) \right). \tag{7}$$

The following equality is the cornerstone of the method and justifies the introduction of (6). If θ satisfies $\mathbb{E}|\theta(Y)K_h(u - \Phi(X)| < \infty$, for all $u \in \Phi(A)$,

$$\mathbb{E}\left[\theta(Y)K_h(u - \Phi(X)\right] = K_h \star \left(g \mathbf{1}_{\Phi(A)}\right)(u) := g_h(u), \tag{8}$$

where $\star$ is the convolution product. It shows that $\hat{g}_h$ is an empirical version of g_h and thus $\hat{s}_h$ in (7) suits well to estimate s. Let us give examples covered by the above framework.

Example 1 (standard random design regression with additive error term): we observe (X_i, Y_i) with $Y_i = s(X_i) + \varepsilon_i$, $(\varepsilon_i)_{i=1,\dots,n}$ is independent of $(X_i)_{i=1,\dots,n}$, $\mathbb{E}[\varepsilon_i^2] < \infty$ and $\mathbb{E}[\varepsilon_i] = 0$. We choose $\Phi(x) = F_X(x)$, the cumulative distribution function (c.d.f. in the sequel) of X and assume that $\Phi : A \to \Phi(A)$ is invertible.

Example 2 (Heteroskedastic model): $Y_i = \sigma(X_i)\varepsilon_i$, $(\varepsilon_i)_{i=1,\dots,n}$ independent of $(X_i)_{i=1,\dots,n}$, $\mathbb{E}[\varepsilon_i^2] = 1$, $\mathbb{E}[\varepsilon_i] = 0$, $\Phi(x) = F_X(x)$, with $\Phi : A \to \Phi(A)$ invertible. Here $s(x) = \sigma^2(x) = \mathbb{E}[Y_i^2 | X_i = x]$.

Example 3 (Interval censoring, Case 1): the observation is (X_i, Y_i) where $Y_i = \mathbf{1}_{Z_i \leq X_i}$, $Z_i, X_i \geq 0$ are independent event occurence times, Y_i indicates whether Z_i (the time of interest) occurs before X_i (the so-called "examination time") or not and Z_i is not observed. The target function is $s(x) = \mathbb{P}(Z_i \leq x) = \mathbb{E}[Y_i | X_i = x]$. We choose $\Phi = F_X$.

Example 4 (Hazard rate estimation from right censored-data): the observation is $X_i = Z_i \wedge C_i$, $Y_i = \mathbf{1}_{Z_i \leq C_i}$, where Z_i and C_i are not observed and independent, $Z_i \geq 0$ is a lifetime and $C_i \geq 0$ is a censoring time. The function s of interest is the hazard rate function $s(x) = f_Z(x)/(1 - F_Z(x))$, where f_Z (resp. F_Z) is the density (resp. the c.d.f.) of Z. This function satisfies

$$s(x) = \frac{f_X(x)}{1 - F_X(x)} \mathbb{E}[Y | X = x]. \tag{9}$$

In this case, we assume $F_X(x) < 1$ for all $x \geq 0$, and take $\Phi(x) = \int_0^x (1 - F_X(t))dt$.

We now define the estimator $\hat{\Phi}$ of the warping function Φ. Instead of estimating Φ with the whole sample $(X_i)_{i=1,\dots,n}$, we assume that another sample $(X_{-i})_{i=1,\dots,n}$, independent of the X_i's, but distributed like them, is available. Thus, we set

$$\hat{\Phi}_n(x) = \begin{cases} \hat{F}_n(x) (\text{Examples 1-3}), \\ \int_0^x \left(1 - \hat{F}_n(t)\right) dt (\text{Example 4}) \text{ where } \hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^n \mathbf{1}_{X_{-i} \leq x}. \end{cases}$$

The introduction of the second sample of variable X is an artefact of the theory: it only allows to avoid dependency problems in the proof of the results, which are technical and cumbersome enough (see the supplementary material Chagny 2013b). This "trick" is also standard in others studies of warped methods (see Kerkyacharian and Picard 2004 or Kulik and Raimondo 2009 *e.g.*). Using a single sample would have required totally different statistic and probabilistic tools. However, we have obviously used only one sample to compute the estimator in the simulation study, see Section 4 (otherwise the comparison with other methods would not have been fair).

Hereafter, we also denote by $\tilde{g}_h$ the pseudo-estimators defined by choosing $\hat{\Phi} = \Phi$ in (6). We coherently set $\tilde{s}_h = \tilde{g}_h \circ \Phi$. They can be used when Φ is known. The theoretical results are the same for $\tilde{s}_h$ and $\hat{s}_h$, up to further technicalities due to the plug-in of an empirical version for Φ. The paper mainly focuses on the proofs of the results for $\tilde{s}$ since they are representative of the statistical and probabilistic tools which are required. Complete proofs are available in the supplementary material Chagny (2013b).

2.2. **Risk of the fixed bandwidth estimator.** In this section, we study the global properties of $\hat{s}_h$ as an estimate of s on A, with a fixed bandwidth h. The quadratic risk weighted by the derivative ϕ of the warping function Φ is the natural criterion in our setting. Let us introduce, for a measurable function t on A,

$$\|t\|_\phi^2 = \int_A t^2(x)\phi(x)dx, \tag{10}$$

and denote by $L^2(A, \phi)$ the space of functions t for which the quantity (10) exists and is finite. We also use the corresponding scalar product $\langle ., . \rangle_\phi$. For t_1, t_2 belonging to $L^2(A, \phi)$, we have

$$\|t_1 \circ \Phi\|_\phi = \|t_1\|_{L^2(\Phi(A))}, \quad \langle t_1 \circ \Phi, t_2 \circ \Phi \rangle_\phi = \langle t_1, t_2 \rangle_{\Phi(A)},$$

where $\|t_1\|^2_{L^2(\Phi(A))} = \int_{\Phi(A)} t_1^2(x)dx$ and $\langle .,. \rangle_{\Phi(A)}$ denotes the usual scalar product on $L^2(\Phi(A))$. Therefore,

$$\|\tilde{s}_h - s\|^2_\phi = \|\tilde{g}_h - g\|^2_{L^2(\Phi(A))}.$$

When K belongs to $L^2(\mathbb{R})$ and $\mathbb{E}[\theta^2(Y_1)] < \infty$, the following bias-variance decomposition of the risk holds (recall that g_h is defined in (8)),

$$\mathbb{E}\left[\|\tilde{s}_h - s\|^2_\phi\right] \leq \|g - g_h\|^2_{L^2(\Phi(A))} + \frac{1}{nh}\mathbb{E}\left[\theta^2(Y_1)\right] \|K\|^2_{L^2(\mathbb{R})}. \tag{11}$$

This inequality holds if $s \in L^2(A, \phi)$, a property which is satisfied if s is bounded on A. This is the case for Examples 1-3, as $\phi = f_X$. In Example 4, we can check that $s \in L^2(A, \phi)$ for all classical distributions for C and Z used in survival analysis (such as exponential, Weibull, Gamma...). The general condition to be checked in this example is $\int_A f_X^2(x)/(1 - F_X(x))dx < \infty$.

The challenge in the general case comes from the plug-in of the empirical $\hat{\Phi}_n$. Though natural, it necessitates lengthy and cumbersome technicalities. This explains why it is not often considered in the literature of warped estimators (see *e.g.* Pham Ngoc 2009 or Chesneau 2007). The following assumptions are required.

(H1') The function s is continuously derivable on A.
(H2') The kernel K is twice continuously derivable, with bounded derivatives K' and K'' on $\mathbb{R}$.
(H3') The set A can be written $A = (0, \tau)$ with finite $\tau > 0$, in Example 4.

Assumption (H1') is somehow restrictive but required for integration by parts (see Section C.3. in Chagny (2013b)). Assumption (H2') permits to use Taylor formulas to deal with the difference $K_h(u - \hat{\Phi}_n(X_i)) - K_h(u - \Phi(X_i))$. This is not a problem as we choose the kernel in practice. We also add (H3'), which is needed to control the difference $\hat{\Phi}_n - \hat{\Phi}$ in Example 4 (see Section B. in Chagny (2013b)). Thanks to technical computations, we obtain the analogous of (11) for $\hat{s}_h$, under mild assumptions. A sketch of the proof is given in Section 5.2. The details are provided in Chagny (2013b).

Proposition 1. *Assume* (H1'), *and* (H2'), *and also* (H3') *for Example 4. If moreover* $h \geq Cn^{-1/5}$ *(*C *a purely numerical constant), there exists* $c > 0$ *such that*

$$\mathbb{E}\left[\|\hat{s}_h - s\|^2_\phi\right] \leq 5\|g - g_h\|^2_{L^2(\Phi(A))} + \frac{c}{nh}. \tag{12}$$

3. Adaptive estimation

As usual, we must choose a bandwidth h which realizes the best compromise between the squared-bias and the variance terms (see (11) and (12)). The choice should be data-driven. For this, we use a method described in Goldenshluger and Lepski (2011), and show that this leads to oracle-type inequalities and adaptive optimal estimators (in the sense of Goldenshluger and Lepski 2011). We begin with the case of known warping function Φ, which permits to develop the theoretical tools in a simpler way and to derive the results with few assumptions and short proofs.

3.1. Case of known Φ.

3.1.1. *Data-driven choice of the bandwidth.* We consider the collection $(\tilde{s}_h)_{h \in \mathcal{H}_n}$, where $\mathcal{H}_n$ is a finite collection of bandwidths, with cardinality depending on n and properties precised below (Assumptions (H2)-(H3)). We introduce the auxiliary estimators, involving two kernels,

$$\tilde{s}_{h,h'}(x) = \tilde{g}_{h,h'}(\Phi(x)) \quad \text{with} \quad \tilde{g}_{h,h'} = K_{h'} \star \left(\tilde{g}_h \mathbf{1}_{\Phi(A)}\right).$$

For a numerical constant $\kappa > 0$ to be precised later on (see Section 4.1 below), we define, for $h \in \mathcal{H}_n$,

$$V(h) = \kappa\left(1 + \|K\|^2_{L^1(\mathbb{R})}\right)\|K\|^2_{L^2(\mathbb{R})}\mathbb{E}\left[\theta^2(Y_1)\right]\frac{1}{nh}. \tag{13}$$

Next, we set

$$A(h) = \max_{h' \in \mathcal{H}_n} \left\{ \|\tilde{s}_{h,h'} - \tilde{s}_{h'}\|_\phi^2 - V(h') \right\}_+, \tag{14}$$

which is an estimation of the squared-bias term (see Lemma 4). Note that $\|\tilde{s}_{h,h'} - \tilde{s}_{h'}\|_\phi^2 = \|\tilde{g}_{h,h'} - \tilde{g}_{h'}\|^2$. Lastly, the adaptive estimator is defined in the following way:

$$\tilde{s} = \tilde{s}_{\tilde{h}} \quad \text{with} \quad \tilde{h} = \arg\min_{h \in \mathcal{H}_n} \{A(h) + V(h)\}. \tag{15}$$

The selected bandwidth $\tilde{h}$ is data-driven. In $V(h)$, the expectation $\mathbb{E}[\theta^2(Y_1)]$ can be replaced by the corresponding empirical mean (see Brunel and Comte 2005, proof of Theorem 3.4 p.465). In Examples 3-4, it can be replaced by 1, its upper-bound.

3.1.2. *Results.* We consider the following assumptions:

(H1) The function s is bounded. Denote by $\|s\|_{L^\infty(A)}$ its sup-norm.

(H2) There exist $\alpha_0 > 0$ and a constant $k_0 \geq 0$ such that $\sum_{h \in \mathcal{H}_n} \frac{1}{h} \leq k_0 n^{\alpha_0}$.

(H3) For all $\kappa_0 > 0$, there exists $C_0 > 0$, such that $\sum_{h \in \mathcal{H}_n} \exp\left(-\frac{\kappa_0}{h}\right) \leq C_0$.

(H4) The kernel K is of order l, *i.e.* for all $j \in \{1, \ldots, l+1\}$, the function $x \mapsto x^j K(x)$ is integrable, and for $1 \leq j \leq l$, $\int_{\mathbb{R}} x^j K(x) dx = 0$.

Assumption (H1) is required to obtain Theorem 2 below. Nevertheless the value $\|s\|_{L^\infty(A)}$ is not needed to compute the estimator (see (15)). This assumption holds in Example 3 ($s \leq 1$ in this case), and in Example 4, for instance when Z has exponential or Gamma distribution. Assumptions (H2)-(H3) mean that the bandwidth collection should not be too large. For instance, the following classical collections satisfy these assumptions:

(1) $\mathcal{H}_{n,1} = \left\{k^{-1},\ k = 1, \ldots, \chi(n)\right\}$ with $\alpha_0 = 2$, $\chi(n) = n$ or $\alpha_0 = 1$, $\chi(n) = \sqrt{n}$.
(2) $\mathcal{H}_{n,2} = \left\{2^{-k},\ k = 1, \ldots, [\ln(n)/\ln(2)]\right\}$, with $\alpha_0 = 1$.

Assumption (H4) is required only to deduce convergence rate from the main nonasymptotic result. We need a moment assumption linked with (H2):

(H5) With α_0 given by (H2), there exists $p > 2\alpha_0$, such that $\mathbb{E}[|\theta(Y) - \mathbb{E}[\theta(Y)|X]|^{2+p}] < \infty$.

If θ is bounded, (H5) evidently holds. In Examples 1 and 2, (H5) is a moment assumption on the noise which is usual in regression settings. It includes *e.g.* the case of Gaussian regression (when the noise ε is a Gaussian variable), under Assumption (H1). Notice also that the smaller α_0, the less restrictive the integrability constraint p on the noise moments.
We prove the following oracle-type inequality:

Theorem 2. *We assume that* (H1)-(H3) *hold in Examples 1-4, and additionally that* (H5) *is fulfilled for Examples 1-2. Then there exist two constants $c_1 > 0$ and $c_2 > 0$, such that:*

$$\mathbb{E}\left[\|\tilde{s} - s\|_\phi^2\right] \leq c_1 \min_{h \in \mathcal{H}_n} \left\{ \|s - s_h\|_\phi^2 + \frac{\mathbb{E}\left[\theta^2(Y_1)\right] \|K\|_{L^2(\mathbb{R})}^2}{nh} \right\} + \frac{c_2}{n}, \tag{16}$$

with $s_h = g_h \circ \Phi$ and $\tilde{s}$ defined by (15). The constant c_1 only depends on $\|K\|_{L^1(\mathbb{R})}$.

The constant c_2 depends on $\|s\|_{L^\infty(A)}$, $\|K\|_{L^1(\mathbb{R})}$ and $\|K\|_{L^2(\mathbb{R})}$ in Examples 3-4, and also on the moment of ε and $\mathbb{E}[s^2(X_1)]$ for Examples 1-2. The adaptive estimator $\tilde{s}$ automatically makes the

squared-bias/variance compromise. The selected bandwidth $\tilde{h}$ is performing as well as the unknown oracle:

$$h^* := \arg\min_{h\in\mathcal{H}_n} \mathbb{E}[\|\tilde{s}_h - s\|_\phi^2].$$

up to the multiplicative constant c_1 and up to a remainding term of order $1/n$, which is negligible. The interest of Inequality (16) is that it is nonasymptotic. Moreover, contrary to usual kernel estimation results, Assumption (H4) is not needed. This is one of the advantages of the bandwidth selection method. Inequality (16) proves that the estimator $\tilde{s}$ is optimal in the oracle sense.

To deduce convergence rates, smoothness classes must be considered to quantify the bias term. Define the Hölder class with order $\beta > 0$ and constant $L > 0$ by

$$\mathcal{H}(\beta, L) = \left\{ t : \mathbb{R} \to \mathbb{R},\ t^{(\lfloor\beta\rfloor)} \text{ exists}, \forall x, x' \in B,\ \left| t^{(\lfloor\beta\rfloor)}(x) - t^{(\lfloor\beta\rfloor)}(x') \right| \leq L|x - x'|^{\beta - \lfloor\beta\rfloor} \right\},$$

where $\lfloor\beta\rfloor$ is the largest integer less than β. We also need the Nikol'skii class of functions:

$$\mathcal{N}_2(\beta, L) = \left\{ t : \mathbb{R} \to \mathbb{R},\ t^{(\lfloor\beta\rfloor)} \text{ exists}, \forall x \in \mathbb{R},\ \int_{\mathbb{R}} \left(t^{(\lfloor\beta\rfloor)}(x' + x) - t^{(\lfloor\beta\rfloor)}(x') \right)^2 dx' \leq L^2 |x|^{2\beta - 2\lfloor\beta\rfloor} \right\}$$

We can now deduce from Theorem 2 the convergence rate of the risk, under regularity assumptions for the auxiliary function g.

Corollary 1. *Let $\bar{g} = g\mathbf{1}_{\phi(A)}$ on $\mathbb{R}$. Assume that*

- *$\bar{g}$ belongs to the Hölder class $\mathcal{H}(\beta, L)$, with $\bar{g}(0) = \tilde{g}(1)$ in Examples 1-3,*
- *$\bar{g}$ belongs to the Nikol'skii class $\mathcal{N}_2(\beta, L)$ in Example 4.*

Assume (H4) *with $l = \lfloor\beta\rfloor$. Then, under the assumptions of Theorem 2,*

$$\mathbb{E}\left[\|\tilde{s} - s\|_\phi^2 \right] \leq C n^{-\frac{2\beta}{2\beta+1}}, \tag{17}$$

where C is a constant which does not depend on n and β.

In Examples 1-3, $\Phi(A) = (0; 1)$ and the Hölder condition is enough. In Example 4, $\Phi(A) = \mathbb{R}_+$ and we need the Nikol'skii condition. Both spaces are standard in kernel estimation, see *e.g.* Tsybakov (2009) and Goldenshluger and Lepski (2011).
We recover the classical optimal rates in nonparametric estimation. Note however that our regularity assumptions are set on g and not s, as long as we do not consider specific warped spaces defined in Kerkyacharian and Picard (2004). The rate (17) is known to be optimal in the minimax sense for the estimation problems we consider (*e.g.*, see Korostelëv and Tsybakov 1993 for regression estimation, Huber and MacGibbon 2004 for hazard rate, and Plancade 2013 for c.d.f. estimation with current-status data), if the two functions have the same regularity parameter.

Remark 1. We have strong conditions on g at the boundary of the support $[0; 1]$, in Examples 1-3. This is nevertheless well-known in kernel estimation, which are rarely "free of boundary effects". This also explains why we restrict the estimation interval for the simulation study, by using the quantiles of the observations X_i (see Section 4). Notice that we may apply recent methods which provide boundary corrections in kernel estimation (see Karunamuni and Alberts 2005 and Bertin and Klutchnikoff 2011 for example), but this is beyond the scope of this paper. Moreover, the comparison against adaptive smoothing splines, which have no such problem of boundary already give good results in practice, see Section 4 below.

3.2. **General case of unknown Φ.** We use the plug-in device of $\hat{\Phi}_n$ in the definition of the quantities (13) and (14) to introduce a criterion like (15) which is fully data-driven in the general case of unknown warped function. To limit the technicalities, we focus on one of the following bandwidth collection $\mathcal{H}_n$: $\mathcal{H}_{n,1} = \{k^{-1},\ k = 1, \ldots, [(n/\ln(n))^{1/5}]\}$ or $\mathcal{H}_{n,2} = \{2^{-k},\ k = 1, \ldots, [\ln(n)/(6\ln(2))]\}$.

The selection of an estimator in the collection $(\hat{s}_h)_{h\in\mathcal{H}_{n,l}}$, $l=1,2$, is done with

$$\hat{h} = \arg\min_{h\in\mathcal{H}_{n,l}} \left\{ \widehat{A}(h) + 2\widehat{V}(h) \right\}, \tag{18}$$

with

$$\widehat{A}(h) = \max_{h'\in\mathcal{H}_n} \left\{ \|\hat{g}_{h,h'} - \hat{g}_{h'}\|^2_{L^2(\Phi(A))} - \widehat{V}(h') \right\}_+, \tag{19}$$

and $\widehat{V}(h) = \kappa' \ln^2(n)/nh$, where, κ' is a purely numerical constant given by the proofs (see Chagny 2013b, Section D) and tuned in practice (see Remark 2). Finally, we define $\hat{s} = \hat{s}_{\hat{h}} \circ \hat{\Phi}_n$. The following result is the equivalent to Theorem 2.

Theorem 3. *We assume that* (H1), (H1')-(H2') *hold in Examples 1-4, and additionally that* (H5) *is fulfilled for Examples 1-2 (with $\alpha_0 = 1$), and* (H3') *for Example 4. Then there exist two constants $c_1', c_2' > 0$, such that, for n large enough,*

$$\mathbb{E}\left[\|\hat{s} - s\|^2_\phi\right] \leq c_1' \min_{h\in\mathcal{H}_{n,l}} \left\{ \|s - s_h\|^2_\phi + \frac{\ln^2(n)}{nh} \right\} + \frac{c_2'}{n}, \tag{20}$$

with $s_h = g_h \circ \Phi$, and for $l = 1, 2$.

The proof of Theorem 3, which follows the same scheme than the one of Theorem 2, is provided in the supplementary material Chagny (2013b). The values of the constants are specified in the proof. They depend on K and s, but neither on n, nor on the bandwidth h. The term $\widehat{V}$ in the definition (18) of the criterion deserves some comments. It can be compared to the "penalty" V defined in the toy case of known Φ in (13). The order of magnitude is not exactly the same: here, we have an extra-logarithmic factor. It is due to the substitution of $\hat{\Phi}_n$ to Φ, and plays a technical role in the proof but it gives what is usually called a nearly optimal bound. Note that it does not depend on $\mathbb{E}[\theta^2(Y_1)]$, contrary to (13).

4. Illustration

To illustrate the procedure, we only focus on two of the four examples: the additive regression (Example 1), and the estimation of c.d.f. under interval censoring case I (Example 3). As we cannot reasonably pretend to compare our method with all the adaptive estimators of the literature, we choose to concentrate on the comparison with adaptive least-squares (LS in the sequel) estimators.

4.1. Implementation of the warped-kernel estimators.

The theoretical study allows the choice of several kernels and bandwidth collections. For practical purpose, we consider the Gaussian kernel, $K : x \mapsto e^{-x^2/2}/\sqrt{2\pi}$, which satisfies Assumption (H4) with $l = 1$. It has the advantage of having simple convolution-products:

$$\forall h, h' > 0, \quad K_h \star K_{h'} = K_{\sqrt{h^2+h'^2}}. \tag{21}$$

The experiment is conducted with the dyadic collection $\mathcal{H}_{n,2}$ defined above. The larger collection $\mathcal{H}_{n,1}$ has also been tested: since it does not really improve the results but increases the computation time, we only keep the other collection. Besides, the simulations are performed in the case of unknown Φ. Therefore in Examples 1 and 3, the estimator is

$$\hat{s} : x \mapsto \frac{1}{n}\sum_{i=1}^n \theta(Y_i) K_{\hat{h}}(\hat{F}_n(x) - \hat{F}_n(X_i)),$$

with $\hat{F}_n$ the empirical c.d.f. of the X_i's. Then, the estimation procedure can be decomposed in some steps:

• Simulate a data sample (X_i, Y_i), $i = 1, \ldots, n$, fitting Example 1 or 3.

• Compute $\widehat{V}(h)$ and $\widehat{A}(h)$ for each $h \in \mathcal{H}_{n,1}$. For $\widehat{V}(h)$, we set $\kappa' = 0.05$ in Example 1, and $\kappa' = 0.3$ in Example 3 (see Remark 2 below). For $\widehat{A}(h)$, thanks to (21), the auxiliary estimates are easily computed: $\hat{s}_{h,h'} = \hat{s}_{\sqrt{h^2+h'^2}}$. The L^2-norm is then approximated by a Riemann sum:

$$\|\hat{g}_{h,h'} - \hat{g}_{h'}\|^2_{L^2(\Phi(A))} \approx \frac{1}{N}\sum_{k=1}^{N}\left(\hat{g}_{h,h'}(u_k) - \hat{g}_{h'}(u_k)\right)^2,$$

where $N = 50$, and $(u_k)_k$ are grid points evenly distributed across $(0;1)$.
• Select $\hat{h}$ such that $\widehat{A}(h) + 2\widehat{V}(h)$ is minimum.
• Compute $\hat{s}_{\hat{h}}$.

Remark 2. The computation of the selection criterion (18) requires a value for κ'. A lower bound for its theoretical value is provided by the proof: it is very pessimistic due to rough upper-bounds (for the sake of clarity), and obviously useless in practice, like in most model selection devices. We have carried out a large number of simulations to tune it, prior to the comparison with the other estimates. We have plotted the quadratic risk in function of κ' for various models and we have chosen one of the first values leading simultaneously to a reasonable risk and a reasonable complexity of the selected bandwidth. Such procedure is classical, and required whatever the chosen selection device (see *e.g.* Sansonnet 2013 for a similar tuning method in wavelet thresholding, or Chagny 2013a for details of tuning in model selection).

4.2. **Example 1: additive regression.** We compare the warped kernel method (WK) with the adaptive estimator studied in Baraud (2002). It is a projection estimator, developed in an orthogonal basis of $L^2(A)$, and built with a penalized least-squares contrast. The experiment is carried out with the Matlab toolbox FY3P, written by Yves Rozenholc, and available on his web page `http://www.math-info.univ-paris5.fr/~rozen/YR/Softwares/Softwares.html`. A regular piecewise polynomial basis is used, with degrees chosen in an adaptive way. Since the kernel we choose has only one vanishing moment, the comparison is fair if we consider polynomials with degrees equal to or less than 1. We denote by LS1 the resulting estimator. However, as shown below, we will see that the warped-kernel generally outperforms the least-square, even if we use polynomials with degree at most 2 (LS2). We also experiment the Fourier basis for the least-squares estimator, but the results are not as good as the polynomial basis. Thus, we do not mention the values of the risks.

The procedure is applied for different regression functions, design and noise. The main goal is to illustrate the sensibility of the estimation to the underlying design distribution. Its influence is explored through four distributions: two classical ones, $\mathcal{U}_{[0;1]}$, the uniform distribution on the interval $[0;1]$, and $\mathcal{N}(0.5, 0.01)$, a Gaussian distribution (with mean 0.5 and variance 0.01); and two distributions which are more original (more irregular), $\gamma(4, 0.08)$, the Gamma distribution, with parameters 4 and 0.08 (0.08 is the scale parameter), and $\mathcal{BN}$ a bimodal Gaussian distribution, with density $x \mapsto c(\exp(-200(x-0.05)^2) + \exp(-200(x-0.95)^2))$ (c is a constant adjusted to obtain a density function).

We focus on the three following regression functions

$$\begin{aligned} s_1 &: x \mapsto x(x-1)(x-0.6) \\ s_2 &: x \mapsto -\exp(-200(x-0.1)^2) - \exp(-200(x-0.9)^2) + 1 \\ s_3 &: x \mapsto \cos(4\pi x) + \exp(-x^2) \end{aligned}$$

The first two functions satisfy the "periodicity" assumptions of Corollary 1, while we choose the function s_3 to show that our procedure also leads to satisfactory results if the assumption does not hold. The choice of the function s_2 combined with the design $\mathcal{BN}$ leads to a data set which presents a "hole" (see Figure 2 (a)).

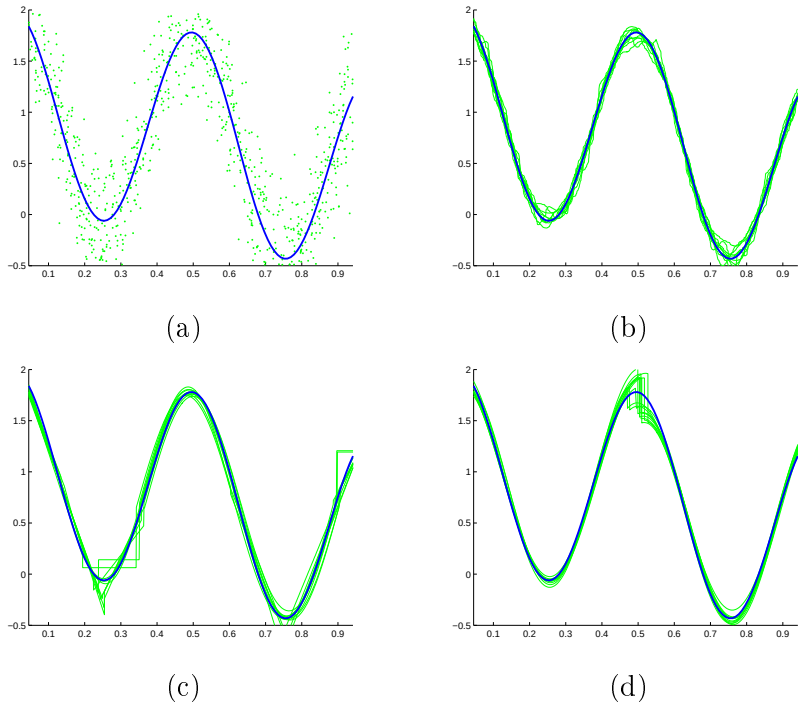


FIGURE 1. Estimation in Example 1, with true regression function s_3, design distribution $\mathcal{U}_{(0;1)}$, and $n = 1000$. (a) points: data $(X_i, Y_i)_i$, thick line: true function s_3. (b)-(c)-(d) beams of 20 estimators built from i.i.d. sample (thin lines) and true function (thick line): warped kernel estimators (subplot (b)), least-squares estimator in piecewise polynomial bases with degree at most 1 (subplot (c)) or 2 (subplot (d)).

We also test the sensibility of the method to the noise distribution: contrary to the underlying design distribution, it does not seem to affect the results. Thus, we present the simulation study for a Gaussian centered noise, with variance σ^2. The value of σ is chosen in such a way that the signal-to-noise ratio (the ratio of the variance of the signal $\mathrm{Var}(s(X_1))$ over the variance of the noise $\mathrm{Var}(\varepsilon_1)$) approximately equals 2.

Beams of estimators (WK, LS1, and LS2) are presented in Figures 1 and 2, with the generated data-sets and the function to estimate. Precisely, Figure 1 shows a regular case: all the methods estimate correctly the signal. Figure 2 depicts the case where a hole occurs in the design density: the estimator built with warped kernel behaves still correctly, even if the data are very inhomogeneous, while the estimator LS1, with which the comparison is fair, failed to detect the hole.

A study of the risk is reported in Tables 1 and 2, for the sample sizes $n = 60, 200, 500$ and 1000. The MISE is obtained by averaging the following approximations of the ISE values, for $j \in \{1, \ldots, J = 200\}$, computed with J sample replications:

$$ISE_j = \frac{b-a}{N} \sum_{k=1}^{N} \left(\tilde{s}(x_k) - s(x_k)\right)^2,$$

where $\tilde{s}$ stands for one of the estimators, b is the quantile of order 95% of the X_i and a is the quantile of order 5%. The $(x_k)_{k=1,\ldots,N}$ are the sample points falling in $[a; b]$. Tables 1 and 2 display the values computed for our method WK, and for the estimators LS1 and/or LS2: for the regression functions s_1 and s_2 (Table 1), the warped-kernel strategy always leads to smaller risk values than

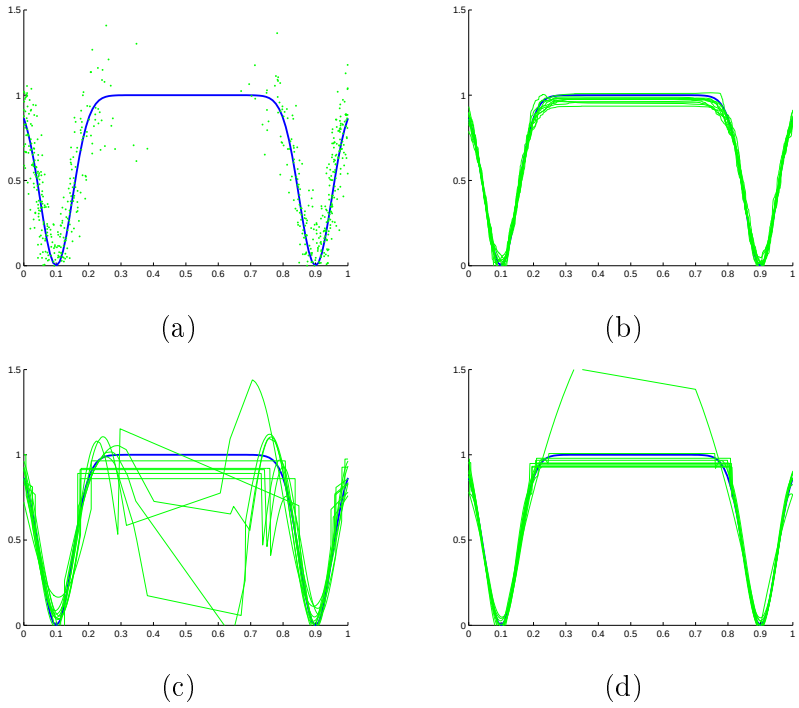


FIGURE 2. Estimation in Example 1, with true regression function s_2, design distribution $\mathcal{BN}$, and $n = 1000$. (a) points: data $(X_i, Y_i)_i$, thick line: true function s_2. (b)-(c)-(d) beams of 20 estimators built from i.i.d. sample (thin lines) and true function (thick line): warped kernel estimators (subplot (b)), least-squares estimator in piecewise polynomial bases with degree at most 1 (subplot (c)) or 2 (subplot (d)).

LS1. Thus, we only mention the risks of the estimator LS2: even if the comparison is quite unfair (following the theoretical results, see the explanations above), the WK and the LS2 estimators are comparable in terms of performance and in 56% of the examples, the risks of the warped-kernel estimator are smaller than the ones of LS2 estimator. For the third function (Table 2), our method still outperforms the least-squares estimate in polynomial basis of degree at most 1, whatever the design distribution is. The comparison with least-squares in polynomial basis with degree at most 2 leads to mixed results even if the values for both estimators are in the same range. Our estimators still behaves correctly (compared to LS1, see also Figure 1), but not as well as for the first two functions. Considering the definition of s_3, the equality $s_3(a) = s_3(b)$, required for the theoretical convergence rate does not hold, which may explain the result.

To conclude, if both methods LS2 and WK lead to comparable risks, it remains that our procedure have some advantages, compared to adaptive least-squares methods. First it is easier to implement, since it does not require any matrix inversion (compared to any LS strategy, see Baraud 2002). Then, keeping in mind that the comparison is fair when choosing piecewise polynomials with degree at most 1, the risk values are always smaller for the warped-kernel estimates, in the studied examples. Finally, we are able to recover a signal even with an irregular design, while the least-squares methods fail in that case.

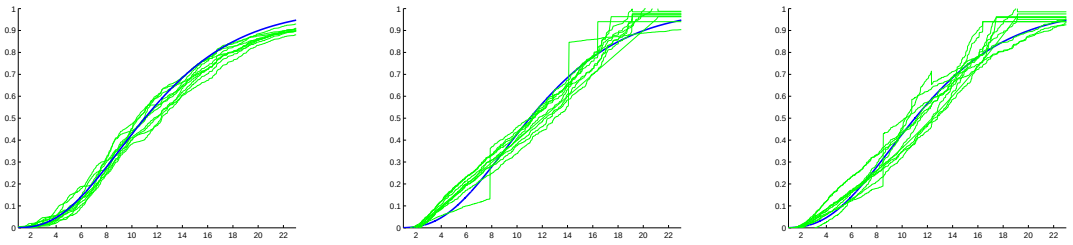
4.3. **Example 3: Interval censoring, case 1.** The same comparison is carried out for the estimation of the c.d.f. under interval censoring. The adaptive least-squares estimate is provided by Brunel and Comte (2009), and the same Matlab toolbox is used for its implementation: recall that

s	X	σ	$n=60$	200	500	1000	Method
s_1	$\mathcal{U}_{[0;1]}$	$\sqrt{.0006}$	0.0889	0.0218	0.0169	0.0167	WK
			0.0856	0.0397	0.0256	0.0229	LS2
	$\gamma(4,0.08)$	5.10^{-5}	0.0052	0.0033	0.0004	0.0003	WK
			0.0097	0.004	0.0017	0.0012	LS2
	$\mathcal{N}(0.5,0.01)$	0.011	0.0049	0.0020	0.0008	0.0005	WK
			0.0020	0.0012	0.0010	0.0008	LS2
	$B\mathcal{N}$	0.022	0.524	0.422	0.267	0.205	WK
			0.166	0.054	0.038	0.029	LS2
s_2	$\mathcal{U}_{[0;1]}$	0.17	16.35	6.791	3.51	0.837	WK
			33.212	2.058	0.691	0.407	LS2
	$\gamma(4,0.08)$	0.08	1.885	0.354	0.204	0.147	WK
			4.047	0.801	0.552	0.429	LS2
	$\mathcal{N}(0.5,0.01)$	0.01	0.0619	0.0186	0.0079	0.0006	WK
			0.0078	0.0014	0.0001	0.0001	LS2
	$B\mathcal{N}$	0.18	12.052	5.279	1.698	1.041	WK
			52.668	11.009	5.817	1.215	LS2

TABLE 1. Values of MISE $\times 1000$ averaged over 200 samples, for the estimators of the regression function (Example 1), built with the warped kernel method (WK) or the least-squares methods, with piecewise polynomials of degree at most 2 (LS2).

s	X	σ	$n=60$	200	500	1000	Method
s_3	$\mathcal{U}_{[0;1]}$	0.35	0.2803	0.1055	0.0463	0.0275	WK
			1.2506	0.4530	0.1261	0.0571	LS1
			0.3107	0.0748	0.0420	0.0332	LS2
	$\gamma(4,0.08)$	0.44	0.1962	0.0628	0.0387	0.0331	WK
			0.4126	0.1334	0.0481	0.0373	LS1
			0.2321	0.0555	0.0206	0.0086	LS2
	$\mathcal{N}(0.5,0.01)$	0.44	0.0634	0.0245	0.0128	0.0861	WK
			0.1045	0.0396	0.0210	0.0108	LS1
			0.0375	0.0139	0.0103	0.0064	LS2
	$B\mathcal{N}$	0.32	0.4438	0.1362	0.0949	0.0793	WK
			1.8253	0.5879	0.1721	0.1232	LS1
			0.6666	0.3038	0.0949	0.0457	LS2

TABLE 2. Values of MISE $\times 10$ averaged over 200 samples, for the estimators of the regression function (Example 1), built with the warped kernel method (WK) or the least-squares methods, with piecewise polynomials of degree at most 1 or 2 (LS1 or LS2).



(a) (b) (c)

FIGURE 3. Estimation in Example 3, in model M7, and $n = 1000$. (a)-(b)-(c) beams of 20 estimators built from i.i.d. sample (thin lines) and true function (thick line): warped kernel estimators (subplot (a)), least-squares estimator in piecewise polynomial bases with degree at most 1 (subplot (b)) or 2 (subplot (c)).

the target function can be seen as a regression function: $s(x) = \mathbb{P}(Z \leq x) = \mathbb{E}[\mathbf{1}_{Z \leq x} | X = x]$. To make the comparison foreseeable, the estimation set A is calibrated as it is done in Brunel and Comte (2009), such that most of the data belong to this interval. Different models are considered for generating the data. We shorten "follows the distribution" by the symbol "$\sim$".

- M1: $X \sim \mathcal{U}_{[0;1]}$, and $Z \sim \mathcal{U}_{[0;1]}$, $A = (0;1)$ (for instance, the target function is $F_Z : x \mapsto x$),
- M2: $X \sim \mathcal{U}_{[0;1]}$, and $Z \sim \chi_2(1)$ (Chi-squared distribution with 1 degree of freedom), $A = (0;1)$,
- M3: $X \sim \mathcal{E}(1)$ (exponential distribution with mean 1), and $Z \sim \chi_2(1)$, $A = (0;1.2)$,
- M4: $X \sim \beta(4,6)$ (Beta distribution of parameter (4,6)), $Z \sim \beta(4,8)$, $A = (0;0.5)$,
- M5: $X \sim \beta(4,6)$, $Z \sim \mathcal{E}(10)$ (exponential distribution with mean 0.1), $A = (0;0.5)$,
- M6: $X \sim \gamma(4, 0.08)$, $Z \sim \mathcal{E}(10)$, $A = (0, 0.5)$,
- M7: $X \sim \mathcal{E}(0.1)$, $Z \sim \gamma(4,3)$, $A = (1;23)$.

All these models allow to investigate thoroughly the sensibility of the method to the distribution of the examination time X, and to the range of the estimation interval. The first two models and the fourth were also used by Brunel and Comte (2009). Since the design is uniform in two of these examples, and also supported by the set $(0;1)$ in the other, we choose to explore what happens when it is not the case: in models M5, M6 and M7, the estimation interval is either smaller (M5-M6) or much larger (M7) than $(0;1)$. Model M3 is choosen to have a design which is not uniform.

Figure 3 shows the smoothness of warped-kernel estimates compared to the reconstruction obtained with least-squares method. The difference between the estimators is also investigated by computing the MISE for the different models. Table 3 reveals that the warped-kernel estimates can advantageously be used as soon as the design X_i has not a uniform distribution: it always outperforms the least-squares estimators in these cases, whatever the estimation support is, and whatever the chosen distributions are. When the design is uniform, the warped-kernel strategy also leads to acceptable results but is a little less interesting (however still simpler to implement than LS methods): since $F_X(x) = x$, one can clearly understand that it is not useful to warp the data. However, recall that F_X is unknown in practice, thus we cannot assure beforehand that the warping of the data is useless.

The practical advantages of our method are thus definitely to permit to deal with various design distributions (even very irregular ones), and thus to be stable to several data sets in different estimation settings (c.d.f. of current status data or regression estimation).

5. PROOFS

Model	X	Z	[a;b]	$n = 60$	200	500	1000	Method
1	$\mathcal{U}_{[0;1]}$	$\mathcal{U}_{[0;1]}$	$[0;1]$	2.41	1.125	0.975	0.533	WK
				0.63	0.111	0.056	0.024	LS2
2	$\mathcal{U}_{[0;1]}$	$\chi_2(1)$	$[0;1]$	1.558	0.804	0.57	0.415	WK
				1.602	0.44	0.244	0.13	LS2
3	$\mathcal{E}(1)$	$\chi_2(1)$	$[0;1.2]$	1.285	0.614	0.243	0.247	WK
				2.385	0.893	0.651	0.365	LS2
4	$\mathcal{B}(4,6)$	$\mathcal{B}(4,8)$	$[0;0.5]$	0.423	0.236	0.09	0.094	WK
				0.449	0.271	0.117	0.105	LS2
5	$\mathcal{B}(4,6)$	$\mathcal{E}(10)$	$[0;0.5]$	0.388	0.229	0.119	0.103	WK
				0.467	0.261	0.13	0.095	LS2
6	$\gamma(4,0.08)$	$\mathcal{E}(10)$	$[0;0.5]$	0.424	0.166	0.102	0.069	WK
				0.698	0.286	0.162	0.095	LS2
7	$\mathcal{E}(0.1)$	$\gamma(4,3)$	$[1;23]$	14.955	5.145	3.973	2.113	WK
				19.825	11.797	9.738	5.898	LS2

TABLE 3. Values of MISE $\times 100$ averaged over 100 samples, for the estimators of the c.d.f. from current status data (Example 3) built with the warped kernel method (WK) or the least-squares methods, with piecewise polynomials of degree at most 1 or 2 (LS1 or LS2).

5.1. **Proof of Inequality (8).** We have:

$$\begin{aligned}\mathbb{E}\left[\theta(Y)K_h(u-\Phi(X))\right] &= \mathbb{E}\left[\mathbb{E}\left[\theta(Y)|X\right]K_h(u-\Phi(X))\right], \\ &= \int_A K_h(u-\Phi(x))\mathbb{E}\left[\theta(Y)|X=x\right]f_X(x)dx.\end{aligned}$$

We set $u' = \Phi(x)$, thus $du' = \phi(x)dx$. Therefore,

$$\begin{aligned}\mathbb{E}\left[\theta(Y)K_h(u-\Phi(X))\right] &= \int_{\Phi(A)} K_h(u-u')\mathbb{E}\left[\theta(Y)|X=\Phi^{-1}(u)\right]f_X\left(\Phi^{-1}(u)\right)\frac{du}{\phi\circ\Phi^{-1}(u)}, \\ &= \int_{\Phi(A)} K_h(u-u')s\circ\Phi^{-1}(u)du.\end{aligned}$$

□

5.2. **Sketch of the proof of Proposition 5.2.** We need to specify the notation. The goal is to study the risk of $\hat{s}_h$, $h \in \mathcal{H}_n$ defined when Φ is unknown. We denote it by $\hat{s}_h^{\hat{\Phi}_n,\hat{\Phi}_n}$. We have

$$\hat{s}_h^{\hat{\Phi}_n,\hat{\Phi}_n} = \hat{g}_h^{\hat{\Phi}_n}\circ\hat{\Phi}_n, \text{ with } \hat{g}_h^{\hat{\Phi}_n}(u) = \frac{1}{n}\sum_{i=1}^n \theta(Y_i)K_h\left(u-\hat{\Phi}_n(X_i)\right).$$

Moreover, $\tilde{s}_h$ is denoted by $\hat{s}_h^{\Phi,\Phi} = \hat{g}_h^{\Phi}\circ\Phi$ with $\hat{g}_h^{\Phi}(u) = (1/n)\sum_{i=1}^n \theta(Y_i)K_h(u-\Phi(X_i)) = \tilde{g}_h(u)$ previously introduced. Coherently, we also set $\hat{s}_h^{\hat{\Phi}_n,\Phi} = \hat{g}_h^{\hat{\Phi}_n}\circ\Phi$. The following decomposition, which

permits to come down to the study of $\tilde{s}$, is the key of the proof: $\left\|\hat{s}_h^{\hat{\Phi}_n,\hat{\Phi}_n} - s\right\|_\phi^2 \leq 5\sum_{l=0}^3 T_l^h$, with

$$\begin{aligned}
T_0^h &= \left\|\hat{s}_h^{\Phi,\Phi} - s_h\right\|_\phi^2 + \left\|s_h^\Phi - s\right\|_\phi^2, \\
T_1^h &= \left\|\hat{s}_h^{\hat{\Phi}_n,\Phi} - \hat{s}_h^{\Phi,\Phi} - \mathbb{E}\left[\hat{s}_h^{\hat{\Phi}_n,\Phi} - \hat{s}_h^{\Phi,\Phi}\,|(X_{-i})\right]\right\|_\phi^2, \\
T_2^h &= \left\|\hat{s}_h^{\hat{\Phi}_n,\hat{\Phi}_n} - \hat{s}_h^{\hat{\Phi}_n,\Phi} - \mathbb{E}\left[\hat{s}_h^{\hat{\Phi}_n,\hat{\Phi}_n} - \hat{s}_h^{\hat{\Phi}_n,\Phi}\,|(X_{-i})\right]\right\|_\phi^2, \\
T_3^h &= \left\|\mathbb{E}\left[\hat{s}_h^{\hat{\Phi}_n,\hat{\Phi}_n} - \hat{s}_h^{\Phi,\Phi}\,|(X_{-i})\right]\right\|_\phi^2,
\end{aligned}$$

where $\mathbb{E}[Z|(X_{-i})]$ is the conditional expectation of a variable Z given the sample $(X_{-i})_{i=1,\dots,n}$. The term T_0^h has been bounded in Inequality (11). For the three others, the challenge is to prove that they are bounded by a quantity with order of magnitude $1/(nh)$.

Notice also that the same splitting is required to prove Theorem 3. All the details are given in Chagny (2013b).

5.3. **Proof of Theorem 2.** The proof is representative of the one of Theorem 3, which is thus deferred to the supplementary Chagny (2013b). Let $h \in \mathcal{H}_n$ be fixed. We start with the following decomposition for the loss of the estimator $\tilde{s} = \tilde{s}_{\tilde{h}}$:

$$\begin{aligned}
\left\|\tilde{s}_{\tilde{h}} - s\right\|_\phi^2 &= \left\|\tilde{g}_{\tilde{h}} - g\right\|_{L^2(\Phi(A))}^2, \\
&\leq 3\left\|\tilde{g}_{\tilde{h}} - \tilde{g}_{h,\tilde{h}}\right\|_{L^2(\Phi(A))}^2 + 3\left\|\tilde{g}_{h,\tilde{h}} - \tilde{g}_h\right\|_{L^2(\Phi(A))}^2 + 3\left\|\tilde{g}_h - g\right\|_{L^2(\Phi(A))}^2.
\end{aligned}$$

The definitions of $A(h)$ and $A(\tilde{h})$ enable us to write, using the definition of $\tilde{h}$,

$$\begin{aligned}
3\left\|\tilde{g}_{\tilde{h}} - \tilde{g}_{h,\tilde{h}}\right\|_{L^2(\Phi(A))}^2 + 3\left\|\tilde{g}_{h,\tilde{h}} - \tilde{g}_h\right\|_{L^2(\Phi(A))}^2 &\leq 3\left(A(h) + V\left(\tilde{h}\right)\right) + 3\left(A\left(\tilde{h}\right) + V(h)\right), \\
&\leq 6\left(A(h) + V(h)\right),
\end{aligned}$$

Besides, applying also (11), we obtain

$$\mathbb{E}\left[\left\|\tilde{s}_{\tilde{h}} - s\right\|_\phi^2\right] \leq 6\mathbb{E}\left[A(h)\right] + 6V(h) + \frac{\mathbb{E}[\theta^2(Y_1)]\|K\|_{L^2(\mathbb{R})}^2}{nh} + 3\|g_h - g\|_{L^2(\Phi(A))}^2. \tag{22}$$

Therefore, the remainding part of the proof follows from the lemma hereafter.

Lemma 4. *Let $h \in \mathcal{H}_n$ be fixed. Under the assumptions of Theorem 2, there exist constants C_1, C_2 such that,*

$$\mathbb{E}\left[A(h)\right] \leq C_1\|g_h - g\|_{L^2(\Phi(A))}^2 + \frac{C_2}{n}, \tag{23}$$

where the constant C_1 only depends on $\|K\|_{L^1(\mathbb{R})}$.

Applying Inequality (23) in (22) implies (16) by taking the infimum over $h \in \mathcal{H}_n$. This ends the proof of Theorem 2.

□

5.4. **Proof of Lemma 4.** To study $A(h)$, we introduce the auxiliary quantities $g_{h,h'} := K_{h'} \star (g_h \mathbf{1}_{\Phi(A)}) = K_{h'} \star ((K_h \star g\mathbf{1}_{\Phi(A)})\mathbf{1}_{\Phi(A)})$, for any $h' \in \mathcal{H}_n$, and we first split

$$\|\tilde{s}_{h,h'} - \tilde{s}_{h'}\|_\phi^2 = \|\tilde{g}_{h,h'} - \tilde{g}_{h'}\|^2_{L^2(\Phi(A))} \leq 3\left(T_a + T_b + \|\tilde{g}_{h'} - g_{h'}\|^2_{L^2(\Phi(A))}\right), \tag{24}$$

where

$$T_a = \|\tilde{g}_{h,h'} - g_{h,h'}\|^2_{L^2(\Phi(A))}, \;\; T_b = \|g_{h,h'} - g_{h'}\|^2_{L^2(\Phi(A))}.$$

The first term can be bounded as follows.

$$\begin{aligned} T_a &\leq \left\|K_h \star \left(\tilde{g}_{h'}\mathbf{1}_{\Phi(A)} - g_{h'}\mathbf{1}_{\Phi(A)}\right)\right\|^2_{L^2(\mathbb{R})}, \\ &\leq \|K\|^2_{L^1(\mathbb{R})} \left\|\tilde{g}_{h'}\mathbf{1}_{\Phi(A)} - g_{h'}\mathbf{1}_{\Phi(A)}\right\|^2_{L^2(\mathbb{R})} = \|K\|^2_{L^1(\mathbb{R})} \|\tilde{g}_{h'} - g_{h'}\|^2_{L^2(\Phi(A))}, \end{aligned}$$

as $\|u \star v\|_{L^2(\mathbb{R})} \leq \|u\|_{L^1(\mathbb{R})}\|v\|_{L^2(\mathbb{R})}$ (Young convolution inequality). In the same way, $T_b \leq \|K_{h'}\|^2_{L^1(\mathbb{R})}\|g_h - g\|^2_{L^2(\Phi(A))}$. Therefore, Decomposition (24) becomes:

$$\|\tilde{s}_{h,h'} - \tilde{s}_{h'}\|_\phi^2 \leq 3\|K\|^2_{L^1(\mathbb{R})}\|g - g_h\|^2_{L^2(\Phi(A))} + 3(1 + \|K\|^2_{L^1(\mathbb{R})}) \|\tilde{g}_{h'} - g_{h'}\|^2_{L^2(\Phi(A))}.$$

Now, we get back to the definition of $A(h)$ given by (14):

$$\begin{aligned} A(h) \leq 3 \;& \|K\|^2_{L^1(\mathbb{R})}\|g - g_h\|^2_{L^2(\Phi(A))} \\ & + 3(1 + \|K\|^2_{L^1(\mathbb{R})}) \max_{h' \in \mathcal{H}_n} \left(\|\tilde{g}_{h'} - g_{h'}\|^2_{L^2(\Phi(A))} - \frac{V(h')}{3(1 + \|K\|^2_{L^1(\mathbb{R})})} \right)_+ . \end{aligned} \tag{25}$$

We can note that $\|\tilde{g}_{h'} - g_{h'}\|_{L^2(\Phi(A))} = \sup_{t \in \bar{S}(0,1)} \langle \tilde{g}_{h'} - g_{h'}, t\rangle_{\Phi(A)}$, with $\bar{S}(0,1)$ a dense countable subset of $\tilde{S}(0,1) = \{t \in L^1(\Phi(A)) \cap L^2(\Phi(A)),\; \|t\|_{L^2(\Phi(A))} = 1\}$ (thanks to the separability of $L^2(\mathbb{R})$, such a set exists. Now,

$$\begin{aligned} \langle \tilde{g}_{h'} - g_{h'}, t\rangle_{\Phi(A)} &= \frac{1}{n}\sum_{i=1}^n \int_{\Phi(A)} \left\{\theta(Y_i)K_{h'}\left(u - \Phi(X_i)\right) - \mathbb{E}\left[\theta(Y_i)K_{h'}\left(u - \Phi(X_i)\right)\right]\right\} t(u)du \\ &= \nu_{n,h'}(t), \end{aligned}$$

where $\nu_{n,h'}$ is an empirical process. Thus, thanks to (25), it remains to bound the deviations of $\sup_{t \in \bar{S}(0,1)} \nu^2_{n,h'}(t)$. First, we have

$$\begin{aligned} &\mathbb{E}\left[\max_{h' \in \mathcal{H}_n} \left(\sup_{t \in \bar{S}(0,1)} \nu^2_{n,h'}(t) - \frac{V(h')}{3(1 + \|K\|^2_{L^1(\mathbb{R})})} \right)_+ \right] \\ &\leq \sum_{h' \in \mathcal{H}_n} \mathbb{E}\left[\left(\sup_{t \in \bar{S}(0,1)} \nu^2_{n,h'}(t) - \frac{V(h')}{3(1 + \|K\|^2_{L^1(\mathbb{R})})} \right)_+ \right]. \end{aligned}$$

Then, the conclusion results from the following lemma:

Lemma 5. *Under the assumptions of Theorem 2, there exists a constant C such that,*

$$\sum_{h \in \mathcal{H}_n} \mathbb{E}\left[\left(\sup_{t \in \bar{S}(0,1)} \nu^2_{n,h}(t) - \tilde{V}(h) \right)_+ \right] \leq \frac{C}{n},$$

with $\tilde{V}(h) = \delta'\|K\|_{L^2(\mathbb{R})}\mathbb{E}[\theta(Y_1)^2]/(nh)$ for a numerical $\delta' > 0$.

We choose the constant κ involved in the definition of V such that $\tilde{V}(h) \leq V(h)(1 + \|K\|^2_{L^1(\mathbb{R})})/3$. Thus, the proof is complete.

□

5.5. **Proof of Lemma 5.** We write the empirical process

$$\nu_{n,h}(t) = \frac{1}{n}\sum_{i=1}^{n}\psi_{t,h}(X_i,Y_i) - \mathbb{E}\left[\psi_{t,h}(X_i,Y_i)\right], \tag{26}$$
$$\text{with } \psi_{t,h}(X_i,Y_i) = \theta(Y_i)\int_{\Phi(A)} K_h\left(u-\Phi(X_i)\right)t(u)du.$$

The guiding idea is to apply the Talagrand Inequality, in its version given in Klein and Rio (2005). We will use the notations used in Lemma 5 of Lacour (2008) (p.812). If θ is bounded, this inequality can be applied. Otherwise, we have to introduce a truncation.

5.5.1. *Example 1.* Recall that $\Phi = F_X$ and $\Phi(A) = [0;1]$. We split the process $\nu_{n,h}$ into three parts, writing $\nu_{n,h} = \nu_{n,h}^{(1)} + \nu_{n,h}^{(2,1)} + \nu_{n,h}^{(2,2)}$, with, for $l = 1,(2,1),(2,2)$,

$$\nu_{n,h}^{(l)} = \frac{1}{n}\sum_{i=1}^{n}\varphi_{t,h}^{(l)}(Z_i) - \mathbb{E}\left[\varphi_{t,h}^{(l)}(Z_i)\right],$$

$Z_i = X_i$ or (X_i,ε_i), and

$$\begin{array}{l}\varphi_{t,h}^{(1)} : x \mapsto s(x)\int_0^1 K_h(u-F_X(x))t(u)du,\\ \varphi_{t,h}^{(2,1)} : (x,\varepsilon) \mapsto \varepsilon\mathbf{1}_{|\varepsilon|\le\kappa_n}\int_0^1 K_h(u-F_X(x))t(u)du,\\ \varphi_{t,h}^{(2,2)} : (x,\varepsilon) \mapsto \varepsilon\mathbf{1}_{|\varepsilon|>\kappa_n}\int_0^1 K_h(u-F_X(x))t(u)du,\end{array}$$

where we define, for a constant c which will be specified below,

$$\kappa_n = c\frac{\sqrt{n}}{\ln(n)}. \tag{27}$$

We apply Talagrand's Inequality to the first two bounded empirical processes, and bound roughly the last one. Thus, we split:

$$\begin{aligned}\sum_{h\in\mathcal{H}_n}\mathbb{E}\left[\left(\sup_{t\in\bar{S}(0,1)}\nu_{n,h}^2(t) - \tilde{V}(h)\right)_+\right] \le\ & 3\sum_{h\in\mathcal{H}_n}\left\{\mathbb{E}\left[\left(\sup_{t\in\bar{S}(0,1)}\left(\nu_{n,h}^{(1)}(t)\right)^2 - \frac{\tilde{V}_1(h)}{3}\right)_+\right]\right.\\ &+\mathbb{E}\left[\left(\sup_{t\in\bar{S}(0,1)}\left(\nu_{n,h}^{(2,1)}(t)\right)^2 - \frac{\tilde{V}_2(h)}{3}\right)_+\right]\\ &\left.+\,\mathbb{E}\left[\sup_{t\in\bar{S}(0,1)}\left(\nu_{n,h}^{(2,2)}(t)\right)^2\right]\right\},\end{aligned} \tag{28}$$

with the decomposition $\tilde{V}(h) = \tilde{V}_1(h) + \tilde{V}_2(h)$, and, denoting by $\delta'' = \delta'/2$,

$$\tilde{V}_1(h) = 3\delta''\frac{\|K\|_{L^2(\mathbb{R})}^2\mathbb{E}\left[s^2(X_1)\right]}{nh}, \text{ and } \tilde{V}_2(h) = 3\delta''\frac{\|K\|_{L^2(\mathbb{R})}^2\mathbb{E}\left[\varepsilon_1^2\right]}{nh}.$$

Actually, recall that we have $\mathbb{E}[\theta^2(Y_1)] = \mathbb{E}[Y_1^2] = \mathbb{E}[s^2(X_1)] + \mathbb{E}[\varepsilon_1^2]$ here.

We now show that each of the three terms of the right hand-side of (28) is upper-bounded by a quantity of order $1/n$. This will end the proof.

• ***First term of (28).***

Let us begin with $\nu_{n,h}^{(1)}$. To do so, we compute $H^{(1)}$, $M^{(1)}$ and $v^{(1)}$, involved in Lemma 5 of Lacour (2008) (p.812).

- For $M^{(1)}$, let $t \in \bar{S}(0,1)$ and $x \in A$ be fixed:

$$\begin{aligned}\left|\varphi_{t,h}^{(1)}(x)\right| &\leq |s(x)| \int_0^1 |K_h(u - F_X(x))t(u)|\, du \leq |s(x)| \|K_h\|_{L^2(\mathbb{R})} \|t\|_{L^2(\Phi(A))}, \\ &= |s(x)| \frac{\|K\|_{L^2(\mathbb{R})}}{\sqrt{h}} \leq \|s\|_{L^\infty(A)} \frac{\|K\|_{L^2(\mathbb{R})}}{\sqrt{h}} := M^{(1)}.\end{aligned}$$

- For $H^{(1)}$, notice that

$$\nu_{n,h}^{(1)}(t) = \langle \hat{d}_h - g_h, t \rangle_{\Phi(A)}, \text{ with } \hat{d}_h = \frac{1}{n} \sum_{i=1}^n s(X_i) K_h\left(. - F_X(X_i)\right).$$

Thus, thanks to the Young convolution inequality, we obtain,

$$\begin{aligned}\mathbb{E}\left[\sup_{t \in \bar{S}(0,1)} \left(\nu_{n,h}^{(1)}(t)\right)^2\right] &= \mathbb{E}\left[\left\|\hat{d}_h - g_h\right\|_{L^2([0;1])}^2\right], \\ &= \int_0^1 \mathrm{Var}\left(\hat{d}_h(u)\right) du, \text{ since } g_h(u) = \mathbb{E}\left[\hat{d}_h(u)\right], \\ &\leq \int_0^1 \frac{1}{n} \mathbb{E}\left[s^2(X_1) K_h^2\left(u - F_X(X_1)\right)\right] du.\end{aligned}$$

Then, we use the same computation as the one done to bound the variance term in the proof of (11), and set $(H^{(1)})^2 = \|K\|_{L^2(\mathbb{R})}^2 \mathbb{E}[s^2(X_1)]/(nh)$.

- For $v^{(1)}$, we also fix $t \in \bar{S}(0,1)$. Hereafter, we set $\check{K}_h(u) = K_h(-u)$. First,

$$\mathrm{Var}\left(\varphi_{t,h}^{(1)}(X_1)\right) \leq \mathbb{E}\left[\left(\varphi_{t,h}^{(1)}(X_1)\right)^2\right] \leq \|s\|_{L^\infty(A)}^2 \mathbb{E}\left[\left(\int_0^1 K_h(u - F_X(X_1))t(u)du\right)^2\right],$$

and the expectation can be written

$$\begin{aligned}\mathbb{E}\left[\left(\int_0^1 K_h(u - F_X(X_1))t(u)du\right)^2\right] &= \mathbb{E}\left[\left(\check{K}_h * \left(t\mathbf{1}_{[0;1]}\right)\right)^2 \left(F_X(X_1)\right)\right], \\ &= \int_0^1 \left(\check{K}_h * \left(t\mathbf{1}_{[0;1]}\right)\right)^2 (u)du \leq \left\|\check{K}_h * \left(t\mathbf{1}_{[0;1]}\right)\right\|_{L^2(\mathbb{R})}^2, \\ &\leq \left\|\check{K}_h\right\|_{L^1(\mathbb{R})}^2 \left\|t\mathbf{1}_{[0;1]}\right\|_{L^2(\mathbb{R})}^2 = \left\|\check{K}_h\right\|_{L^1(\mathbb{R})}^2 \|t\|_{L^2([0;1])}^2,\end{aligned}$$

thanks to the Young convolution inequality. Therefore,

$$\mathrm{Var}\left(\varphi_{t,h}^{(1)}(X_1)\right) \leq \|s\|_{L^\infty(A)} \|K\|_{L^1(\mathbb{R})}^2 := v^{(1)}.$$

Then, Talagrand's Inequality gives, for $\delta > 0$,

$$\mathbb{E}\left[\left(\sup_{t \in \bar{S}(0,1)} \left(\nu_{n,h}^{(1)}(t)\right)^2 - 2(1+2\delta)\left(H^{(1)}\right)^2\right)_+\right] \leq k_1 \left\{\frac{1}{n} \exp\left(-k_2 \frac{1}{h}\right) + \frac{1}{n^2 h} \exp\left(-k_3 \sqrt{n}\right)\right\},$$

where k_1, k_2, k_3 are three constants which depend on $\mathbb{E}[s^2(X_1)]$, $\|s\|_{L^\infty(A)}$, $\|K\|_{L^1(\mathbb{R})}$ and $\|K\|_{L^2(\mathbb{R})}$. Assumptions (H2)-(H3) lead to

$$\sum_{h \in \mathcal{H}_n} \mathbb{E}\left[\left(\sup_{t \in \bar{S}(0,1)} \left(\nu_{n,h}^{(1)}(t)\right)^2 - 2(1+2\delta)\|K\|_{L^2(\mathbb{R})}^2 \mathbb{E}[s^2(X_1)] \frac{1}{nh}\right)_+\right] \leq \frac{C}{n},$$

with C a constant (which also depends on the previous quantities).

- ***Second term of (28).***

For the second empirical process $\nu_{n,h}^{(2,1)}$, the sketch of the proof is the same: similarly, we compute the quantities involved in the Talagrand Inequality,

$$M^{(2)} = \kappa_n \|K\|_{L^2(\mathbb{R})} \frac{1}{\sqrt{h}}, \quad H^{(2)} = \|K\|_{L^2(\mathbb{R})} \left(\mathbb{E}[\varepsilon_1^2]\right)^{1/2} \frac{1}{\sqrt{nh}}, \quad v^{(2)} = \|K\|_{L^1(\mathbb{R})}^2 \mathbb{E}[\varepsilon_1^2],$$

and we obtain, by Lemma 5 of Lacour (2008) (p.812), for $\delta > 0$,

$$\mathbb{E}\left[\left(\sup_{t\in\bar{S}(0,1)} \left(\nu_{n,h}^{(2,1)}(t)\right)^2 - 2(1+2\delta)\left(H^{(2)}\right)^2\right)_+\right] \leq k_1 \left\{\frac{1}{n}\exp\left(-k_2\frac{1}{h}\right) + \frac{\kappa_n^2}{n^2 h}\exp\left(-k_3\frac{\sqrt{n}}{\kappa_n}\right)\right\},$$

where k_1, k_2, k_3 are three constants which depend on $\mathbb{E}[\varepsilon_1^2]$, $\|K\|_{L^1(\mathbb{R})}$ and $\|K\|_{L^2(\mathbb{R})}$. The first term of the right hand-side is like above. With the definition (27) of κ_n, the sum over $h \in \mathcal{H}_n$ of the second term of the upper bound can be written

$$\sum_{h\in\mathcal{H}_n} \frac{\kappa_n^2}{n^2 h}\exp\left(-k_3\frac{\sqrt{n}}{\kappa_n}\right) = \frac{c^2}{n^{1+k_3/c}\ln^2(n)} \sum_{h\in\mathcal{H}_n} \frac{1}{h}.$$

Consequently, using Assumptions (H2)-(H3) and choosing c in the definition of κ_n such that $c \leq k_3/\alpha_0$, we also obtain for a constant C,

$$\sum_{h\in\mathcal{H}_n} \mathbb{E}\left[\left(\sup_{t\in\bar{S}(0,1)} \left(\nu_{n,h}^{(2,1)}(t)\right)^2 - 2(1+2\delta)\|K\|_{L^2(\mathbb{R})}^2 \mathbb{E}[\varepsilon_1^2]\frac{1}{nh}\right)_+\right] \leq \frac{C}{n}.$$

- ***Third term of (28).***

The last empirical process is $\nu_{n,h}^{(2,2)}(t) = \int_0^1 t(u)\psi(u)du$, with

$$\psi(u) = \frac{1}{n}\sum_{i=1}^n \varepsilon_i \mathbf{1}_{\{|\varepsilon_i|>\kappa_n\}} K_h\left(u - F_X(X_i)\right) - \mathbb{E}\left[\varepsilon_i \mathbf{1}_{\{|\varepsilon_i|>\kappa_n\}} K_h\left(u - F_X(X_i)\right)\right].$$

It is not bounded. Nevertheless, we use the Cauchy-Schwarz Inequality, and the equality $\|t\|_{L^2(\Phi(A))} = 1$, for $t \in \bar{S}(0,1)$

$$\begin{aligned}
\mathbb{E}\left[\sup_{t\in\tilde{S}(0,1)} \left(\nu_{n,h}^{(2,2)}(t)\right)^2\right] &\leq \mathbb{E}\left[\int_0^1 \psi^2(u)du\right], \\
&\leq \frac{1}{n}\mathbb{E}\left[\varepsilon_1^2 \mathbf{1}_{\{|\varepsilon_1|>\kappa_n\}}\right] \mathbb{E}\left[\int_0^1 K_h^2\left(u - F_X(X_1)\right)du\right], \\
&\leq \frac{\|K\|_{L^2(\mathbb{R})}^2}{nh}\mathbb{E}\left[\varepsilon_1^2 \mathbf{1}_{\{|\varepsilon_1|>\kappa_n\}}\right] \leq \frac{\|K\|_{L^2(\mathbb{R})}^2 \kappa_n^{-p}}{nh}\mathbb{E}\left[\varepsilon_1^{2+p}\right].
\end{aligned}$$

Thus, there exists a constant k_1 which depends on $\|K\|_{L^2(\mathbb{R})}$ and $\mathbb{E}[\varepsilon_1^{2+p}]$,

$$\sum_{h\in\mathcal{H}_n} \mathbb{E}\left[\sup_{t\in\bar{S}(0,1)} \left(\nu_{n,h}^{(2,2)}(t)\right)^2\right] \leq k_1 \frac{\kappa_n^{-p}}{n}\sum_{h\in\mathcal{H}_n}\frac{1}{h} = c_1\kappa^{-p}\frac{\ln^p(n)}{n^{1+p/2}}\sum_{h\in\mathcal{H}_n}\frac{1}{h}.$$

The conclusion comes from Assumptions (H2)-(H3), and the choice of $p \geq 2\alpha_0$.

□

5.5.2. *Examples 2-4.* For the multiplicative regression model (Example 2), we split the process into two terms: $\nu_{n,h} = \nu_{n,h}^{(1)} + \nu_{n,h}^{(2)}$, with

$$\begin{aligned}
\nu_{n,h}^{(1)}(t) &= \frac{1}{n}\sum_{i=1}^{n}\left\{\sigma^2(X_i)\varepsilon_i^2\mathbf{1}_{\{|\varepsilon_i|\leq\kappa_n\}}\int_0^1 K_h\left(u-F_X(X_i)\right)t(u)du\right. \\
&\quad \left.-\mathbb{E}\left[\sigma^2(X_i)\varepsilon_i^2\mathbf{1}_{\{|\varepsilon_i|\leq\kappa_n\}}\int_0^1 K_h\left(u-F_X(X_i)\right)t(u)du\right]\right\}, \\
\nu_{n,h}^{(2)}(t) &= \frac{1}{n}\sum_{i=1}^{n}\left\{\sigma^2(X_i)\varepsilon_i^2\mathbf{1}_{\{|\varepsilon_i|>\kappa_n\}}\int_0^1 K_h\left(u-F_X(X_i)\right)t(u)du\right. \\
&\quad \left.-\mathbb{E}\left[\sigma^2(X_i)\varepsilon_i^2\mathbf{1}_{\{|\varepsilon_i|>\kappa_n\}}\int_0^1 K_h\left(u-F_X(X_i)\right)t(u)du\right]\right\},
\end{aligned}$$

where κ_n is still a constant for the proof, which equals $\sqrt{c\frac{\sqrt{n}}{\ln(n)}}$ and $c>0$ is obtained by the computations, like in Example 1. We exactly recover the framework of this previous example: the deviations of the process $\nu_{n,h}^{(1)}$ are bounded thanks to Talagrand's Inequality of Lemma 5 of Lacour (2008) (p.812), and the second one is bounded in the same way as the process $\nu_{n,h}^{(2,2)}$ of the additive regression setting.

For Examples 3-4, there is no point in splitting the process (26), since it is already bounded (recall that $\theta(Y_1)$ is bounded by 1). Thus, we apply the concentration inequality.

Recall that $\Phi(A)=\mathbb{R}_+$. In both of these cases, the quantity M_1 involved in the assumptions of the Talagrand Inequality (Lemma 5 of Lacour 2008, p.812) equals $M_1 = \|K\|_{L^2(\mathbb{R})}/\sqrt{h}$. Moreover, H^2 can be chosen as the upper-bound of the variance term of the estimator $\tilde{g}_h$, that is $H^2 = \|K\|_{L^2(\mathbb{R})}/nh$. Finally, v equals $\|K\|_{L^1(\mathbb{R})}$ for Example 3, and $\|g\|_{L^\infty(\mathbb{R}_+)}\|K\|_{L^1(\mathbb{R})}$ for Example 4.

As an example, let us detail the computation of v in Example 4. Recall that $X = C\wedge Z$, $Y = \mathbf{1}_{Z\leq C}$, s is the hazard rate, and the warping Φ is the function $x\mapsto\int_0^x(1-F_X(t))dt$. Thus, denoting by f_C (respectively f_Z) a density of the variable C (respectively Z), and F_C (respectively F_Z) its c.d.f.,

$$\begin{aligned}
\mathrm{Var}\left(\psi_{t,h}(X_1,Y_1)\right) &\leq \mathbb{E}\left[\left(\psi_{t,h}(X_1,Y_1)\right)^2\right] = \mathbb{E}\left[Y_1\left(\int_{\mathbb{R}_+}K_h(u'-\Phi(X_1))t(u')du'\right)^2\right], \\
&= \int_{\mathbb{R}_+\times\mathbb{R}}\mathbf{1}_{z\leq c}\left(\int_{\mathbb{R}_+}K_h(u'-\Phi(z))t(u')du'\right)^2 f_C(c)f_Z(z)dzdc, \\
&= \int_{\mathbb{R}_+}\left(\int_{\mathbb{R}_+}K_h(u'-\Phi(z))t(u')du'\right)^2 f_Z(z)\left(1-F_C\right)(z)dz.
\end{aligned}$$

We set $z=\Phi^{-1}(u)$. The integral becomes

$$\begin{aligned}
&\int_{\mathbb{R}_+}\left(\int_{\mathbb{R}_+}K_h(u'-\Phi(z))t(u')du'\right)^2 f_Z(z)\left(1-F_C\right)(z)dz \\
&=\int_{\mathbb{R}_+}\left(\int_{\mathbb{R}_+}K_h(u'-u)t(u')du'\right)^2 f_Z\circ\Phi^{-1}(u)\left(1-F_C\right)\circ\Phi^{-1}(u)\frac{du}{\left((1-F_X)\circ\Phi^{-1}(u)\right)}.
\end{aligned}$$

Thanks to the same arguments as the ones used to prove (8) in Section 5.1, we obtain:

$$\begin{aligned}\mathrm{Var}\left(\varphi_{t,h}(X_1,Y_1)\right) &\leq \int_{\mathbb{R}_+} g(u)\left(\int_{\mathbb{R}_+} K_h(u'-u)t(u')du'\right)^2 du,\\ &= \int_{\mathbb{R}_+} g(u)\left(K_h * \left(t\mathbf{1}_{\mathbb{R}_+}\right)(u)\right)^2 du \leq \|g\|_{L^\infty(\mathbb{R}_+)}\left\|\check{K}_h * \left(t\mathbf{1}_{\mathbb{R}_+}\right)\right\|_{L^2(\mathbb{R})},\\ &\leq \|g\|_{L^\infty(\mathbb{R}_+)}\left\|\check{K}_h\right\|_{L^1(\mathbb{R})}\left\|\left(t\mathbf{1}_{\mathbb{R}_+}\right)\right\|_{L^2(\mathbb{R})} = \|g\|_{L^\infty(\mathbb{R}_+)}\|K\|_{L^1(\mathbb{R})} := v.\end{aligned}$$

Once we have the three quantities, we easily apply the Talagrand Inequality and the proof is complete by using Assumptions (H2)-(H3), like above (see the computations in Example 1).

□

5.6. **Proof of Corollary 1.** We must bound the bias term of the right hand-side of Inequality (16) (Theorem 2). Actually, if we prove that

$$\|s - s_h\|_\phi^2 \leq Ch^{2\beta},$$

where C is a constant, then the proof of the Corollary will be completed by computing the minimum which is involved in (16). By definition,

$$\|s - s_h\|_\phi^2 = \|g - g_h\|^2_{L^2(\Phi(A))} = \int_{\Phi(A)} (g_h(u) - g(u))^2\, du.$$

We distinguish two cases in the sequel, depending on the considered examples.

Then we distinguish two cases:

5.6.1. *Examples 1-3.* Here, $\Phi(A) = (0;1)$. We start with the definition of g_h: for $u \in \Phi(A)$,

$$\begin{aligned}g_h(u) &= \frac{1}{h}\int_0^1 g(u')K\left(\frac{u-u'}{h}\right)du' = \int_{\frac{u-1}{h}}^{\frac{u}{h}} g(u-hz)K(z)dz,\\ &= \int_{\frac{u-1}{h}}^{\frac{u}{h}} \bar{g}(u-hz)K(z)dz = \int_{\mathbb{R}} \bar{g}(u-hz)K(z)dz.\end{aligned}$$

Thus, since $\int_{\mathbb{R}} K(u)du = 1$,

$$\bar{g}_h(u) - g(u) = \int_{\mathbb{R}} K(z)\bar{g}(u-hz)dz - \bar{g}(u) = \int_{\mathbb{R}} K(z)\left[\bar{g}(u-hz) - \bar{g}(u)\right]dz. \tag{29}$$

We use a Taylor-Lagrange formula for $\bar{g}$: for $u \in (0;1)$, and $z \in \mathbb{R}$, there exists $\theta \in (0;1)$ such that

$$\bar{g}(u-hz) - g(u) = -hz\bar{g}'(u) + \frac{(-hz)^2}{2!}\bar{g}''(u) + \cdots + \frac{(-hz)^{l-1}}{(l-1)!}\bar{g}^{(l-1)}(u) + \frac{(-hz)^l}{l!}\bar{g}^{(l)}(u-\theta hz),$$

with $l = \lfloor\beta\rfloor$. With Assumption (K_l), we obtain

$$\|s - s_h\|_\phi^2 \leq \left(\int_{z\in\mathbb{R}} |K(z)|\frac{|hz|^l}{l!}\left\{\int_{u=0}^1 \left\{\bar{g}^{(l)}(u-\theta hz) - \bar{g}^{(l)}(u)\right\}^2 du\right\}^{1/2} dz\right)^2.$$

Since $\bar{g}$ belongs to the Hölder space $\mathcal{H}(\beta, L)$,

$$\begin{aligned}\left[\int_{u=0}^1 \left\{\bar{g}^{(l)}(u-\theta hz) - \bar{g}^{(l)}(u)\right\}^2 du\right]^{1/2} &\leq \left[\int_{u=0}^1 L^2(\theta hu)^{2(\beta-l)}du\right]^{1/2},\\ &= L|hz|^{\beta-l},\end{aligned}$$

which enables us to conclude.

□

5.6.2. *Example 4.* Here, $\Phi(A) = \mathbb{R}_+$. Similarly, we first obtain Equality (29). Then, the idea is the same as in Examples 1-3, but since we integrate over an unbounded subset, we choose an integrated remainding term in the Taylor formula:

$$\bar{g}(u-hz)-\bar{g}(u) = -hz\bar{g}'(u)+\frac{(-hz)^2}{2!}\bar{g}''(u)+\cdots+\frac{(-hz)^{l-1}}{(l-1)!}\bar{g}^{(l-1)}(u)+\frac{(-hz)^l}{(l-1)!}\int_0^1(1-\theta)^{l-1}\bar{g}^{(l)}(u-\theta hz)d\theta.$$

The reasoning is then the same as in density estimation (see Tsybakov 2009 for details).

□

Acknowledgements. I would like to thank Fabienne Comte for a wealth of smart advice and carefully readings along this work. I am also grateful to Valentine Genon-Catalot for proofreading earlier version of this paper. Finally, I gratefully acknowledge the referees for carefully reading the manuscript and for numerous suggestions that improved the paper.

Supporting information. Additional information for this article is available online, containing detailed proofs mainly for Proposition 1 and Theorem 3.

References

Akakpo, N. and Durot, C. (2010). Histogram selection for possibly censored data. *Math. Methods Statist.* **19**, 189–218.

Akritas, M. G. (2005). Reverse windows in nonparametric regression. *J. Statist. Res.* **39**, 77–96.

Baraud, Y. (2002). Model selection for regression on a random design. *ESAIM Probab. Statist.* **6**, 127–146 (electronic).

Bertin, K. and Klutchnikoff, N. (2011). Minimax properties of beta kernel estimators. *J. Statist. Plann. Inference* **141**, 2287–2297.

Brunel, E. and Comte, F. (2005). Penalized contrast estimation of density and hazard rate with censored data. *Sankhyā* **67**, 441–475.

Brunel, E. and Comte, F. (2008). Adaptive estimation of hazard rate with censored data. *Comm. Statist. Theory Methods* **37**, 1284–1305.

Brunel, E. and Comte, F. (2009). Cumulative distribution function estimation under interval censoring case 1. *Electron. J. Stat.* **3**, 1–24.

Chagny, G. (2013a). Penalization versus Goldenshluger-Lepski strategies in warped bases regression. *ESAIM Probab. Statist.* **17**, 328–358 (electronic).

Chagny, G. (2013b). Supplementary material for adaptive warped kernel estimator. Technical report.

Chagny, G. (2013c). Warped bases for conditional density estimation. *Math. Methods Statist.* **22**, 253–282.

Chesneau, C. (2007). A maxiset approach of a Gaussian noise model. *TEST* **16**, 523–546.

Chesneau, C. and Willer, T. (2012). Estimation of a cumulative distribution function under interval censoring "case 1" via warped wavelets. *Submitted, hal-00715260, v3* .

Goldenshluger, A. and Lepski, O. (2011). Bandwidth selection in kernel density estimation: oracle inequalities and adaptive minimax optimality. *Ann. Statist.* **39**, 1608–1632.

Huber, C. and MacGibbon, B. (2004). Lower bounds for estimating a hazard. In *Advances in survival analysis*, vol. 23 of *Handbook of Statist.*, pp. 209–226. Elsevier, Amsterdam.

Jewell, N. P. and van der Laan, M. (2004). Current status data: review, recent developments and open problems. In *Advances in survival analysis*, vol. 23 of *Handbook of Statist.*, pp. 625–642. Elsevier, Amsterdam.

Karunamuni, R. J. and Alberts, T. (2005). On boundary correction in kernel density estimation. *Stat. Methodol.* **2**, 191–212.

Kerkyacharian, G. and Picard, D. (2004). Regression in random design and warped wavelets. *Bernoulli* **10**, 1053–1105.

Klein, T. and Rio, E. (2005). Concentration around the mean for maxima of empirical processes. *Ann. Probab.* **33**, 1060–1077.

Korostelëv, A. P. and Tsybakov, A. B. (1993). *Minimax theory of image reconstruction*, vol. 82 of *Lecture Notes in Statistics*. Springer-Verlag, New York.

Kulik, R. and Raimondo, M. (2009). Wavelet regression in random design with heteroscedastic dependent errors. *Ann. Statist.* **37**, 3396–3430.

Lacour, C. (2008). Adaptive estimation of the transition density of a particular hidden Markov chain. *J. Multivariate Anal.* **99**, 787–814.

Ma, S. and Kosorok, M. R. (2006). Adaptive penalized M-estimation with current status data. *Ann. Inst. Statist. Math.* **58**, 511–526.

Mammen, E., Rothe, C. and Schienle, M. (2012). Nonparametric regression with nonparametrically generated covariates. *Ann. Statist.* **40**, 1132–1170.

Mehra, K. L., Ramakrishnaiah, Y. S. and Sashikala, P. (2000). Laws of iterated logarithm and related asymptotics for estimators of conditional density and mode. *Ann. Inst. Statist. Math.* **52**, 630–645.

Müller, H.-G. and Wang, J.-L. (1994). Hazard rate estimation under random censoring with varying kernels and bandwidths. *Biometrics* **50**, 61–76.

Nadaraya, E. (1964). On estimating regression. *Theory of Probability and its Application* **9**, 141–142.

Patil, P. N. (1993). Bandwidth choice for nonparametric hazard rate estimation. *J. Statist. Plann. Inference* **35**, 15–30.

Penskaya, M. (1995). Mean square consistent estimation of a ratio. *Scand. J. Statist.* **22**, 129–137.

Pham Ngoc, T. M. (2009). Regression in random design and Bayesian warped wavelets estimators. *Electron. J. Stat.* **3**, 1084–1112.

Plancade, S. (2013). Adaptive estimation of the conditional cumulative distribution function from current status data. *J. Statist. Plann. Inference* **143**, 1466–1485.

Reynaud-Bouret, P. (2006). Penalized projection estimators of the Aalen multiplicative intensity. *Bernoulli* **12**, 633–661.

Sansonnet, L. (2013). Wavelet thresholding estimation in a poissonian interactions model with application to genomic data. *Scand. J. Statist.* to appear (available online).

Stute, W. (1984). Asymptotic normality of nearest neighbor regression function estimates. *Ann. Statist.* **12**, 917–926.

Stute, W. (1986). Conditional empirical processes. *Ann. Statist.* **14**, 638–647.

Tanner, M. A. and Wong, W. H. (1983). The estimation of the hazard function from randomly censored data by the kernel method. *Ann. Statist.* **11**, 989–993.

Tsybakov, A. B. (2009). *Introduction to nonparametric estimation*. Springer Series in Statistics. Springer, New York. Revised and extended from the 2004 French original, Translated by Vladimir Zaiats.

van de Geer, S. (1993). Hellinger-consistency of certain nonparametric maximum likelihood estimators. *Ann. Statist.* **21**, 14–44.

Watson, G. S. (1964). Smooth regression analysis. *Sankhyā Ser. A.* **26**, 359–372.

Yang, S.-S. (1981). Linear functions of concomitants of order statistics with application to nonparametric estimation of a regression function. *J. Amer. Statist. Assoc.* **76**, 658–662.